DOCUMENT RESUME

ED 139 830 TM 006 323

AUTHOR Littlefield, John H.; And Others
TITLE Assessing the Generalizability of Clinical Rating Scales.
PUB DATE [Apr 77]
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Analysis of Variance; Clinical Experience; *College Teachers; Data Collection; Dental Schools; Dentists; *Generalization; Higher Education; Mathematical Models; Measurement Techniques; *Rating Scales; Research Methodology; *Student Evaluation; *Test Reliability; Test Results; True Scores
IDENTIFIERS Generalizability Theory; Interrater Reliability

ABSTRACT

Generalizability theory extends previous methods of estimating the reliability of rating instruments such that one can estimate the precision of a measurement system for differentiating among students, scales, or any other important dimension. In this study, generalizability theory is applied to faculty ratings of junior and senior dental students on five broadly defined scales. Utilizing two observers, the measurement system differentiates accurately among students, however, interpretation of inter-scale differences for a given student are less dependable. The methodology demonstrated is recommended for use in pilot testing rating instruments for evaluating clinical performance. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED139830

TM006 323

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Assessing the Generalizability of Clinical Rating Scales

John H. Littlefield, Ph.D.
A. J. Murrey, D.D.S.
Russell E. Garman, M.Crim.

University of Texas Health Science Center at San Antonio

Paper Presented at the Annual Meeting of the American Educational Research Association
New York, New York
April 5, 1977

Assessing the Generalizability of Clinical Rating Scales (VAR)

JOHN H. LITTLEFIELD, Ph.D.

A.J. MURREY, D.D.S.

RUSSELL E. GARMAN, M.CRIM.

University of Texas Health Science Center at San Antonio

Generalizability theory extends previous methods of estimating the reliability of rating instruments such that one can estimate the precision of a measurement system for differentiating among students, scales, or any other important dimension. In this study, generalizability theory is applied to faculty ratings of junior and senior dental students on 5 broadly defined scales. Utilizing 2 observers, the measurement system differentiates accurately among students ($p^2 = .76$); however, interpretations of inter-scale differences for a given student are less dependable ($p^2 = .49$). The methodology demonstrated is recommended for use in pilot testing rating instruments for evaluating clinical performance.

Assessing the Generalizability of Clinical Rating Scales

Introduction

Rating scales are a widely used method of collecting data for evaluating clinical performance in the health sciences. Reports of research using rating scales often fail to establish the inter-rater reliability of the measurement system (Rowley, 1976). This is particularly distressing when one observes that techniques for estimating inter-rater reliability of rating instruments have been available for over 25 years (Ebel, 1951). Generalizability theory (Cronbach et. al., 1972; Cardinet et. al., 1976) extends Ebel's work such that one can establish the precision of a measurement system for differentiating among students, objectives or levels on other dimensions. In this study, generalizability theory is applied to demonstrate the consistency with which a particular measurement system can: (a) differentiate among students using five given scales; (b) differentiate among students using only the mean score across scales; and (c) assess inter-scale differences for a given student.

Background

Classical test theory is based upon the concept that observed scores are estimates of a hypothetical parameter called a "true score" (Gulliksen, 1950). The true score for a given individual is defined as the mean score obtained using an infinitely large number of tests. Each test is assumed to be a parallel form of all other tests and all measurement errors (i.e., variability not caused by differences in the trait being measured) are assumed to be random. In constructing and using clinical rating scales, one quickly recognizes that many nonrandom measurement errors exist. For example, individual raters often vary in the strictness with which they apply the rating criteria.

Generalizability theory (Cronbach et. al., 1972) is based upon the concept of a "universe score", namely the mean of all measurements that could have been obtained under a specified set of conditions. In contrast with classical test theory, a measurement procedure is not classified as producing reliable or unreliable scores. Instead, one can generalize to different degrees from one observed score to a particular universe score. Measurement errors are assumed to be affected by systematic sources of variability such as multiple observers. The theory offers a statistical model for assessing the impact of various types of measurement error upon the precision of scores obtained under specific conditions. When the measurement conditions are varied, a separate generalizability coefficient can be computed for each new condition.

In classical test theory, consistency of measurement is defined as the ratio of true differences among students over observed differences. Algebraically, this ratio is defined as:

$$\frac{V_t}{V_t + V_e} \qquad (1)$$

As noted earlier, measurement errors (V_e) are assumed to be random and therefore cannot be measured directly. Generalizability theory, on the other hand, seeks to identify the systematic components of measurement error. For ratings by multiple observers, the ratio above can be redefined as:

$$\frac{V_t}{V_t + (V_r + V_e')} \qquad (2)$$

In formula (2) error variance (V_e) has been partitioned into variance due to multiple raters (V_r) plus unexplained variance (V_e'). The computational

formulas presented later are simply an extension of the principle demonstrated by formula (2). The numerators are an estimate of true variance while the denominators include the true variance plus various sources of error which reduce measurement precision for particular uses. As the denominator increases in size relative to the numerator, the generalizability coefficient will be reduced. A small coefficient indicates that error variance in the measurement system is large relative to observed variation on the dimension which we are attempting to measure. Therefore, we cannot be confident that a given rating is highly representative of the universe score.

Analysis of variance techniques can be used to estimate each variance component in formula (2). The resulting ratio is equivalent to an intraclass correlation coefficient (Ebel, 1951). Interpretations of the numerical size of a generalizability coefficient are analogous to the traditional reliability coefficient (e.g., .80 is considered adequate for most instructional applications).

More recently, Cardinet et. al. (1976) have expanded the theory presented by Cronbach et. al. (1972) to include assessing the generalizability of measurement data for purposes other than differentiating among students (e.g., interpreting inter-scale differences). An application suggested by Cardinet et. al. is included in the examples presented below.

Data Collection and Analysis

A rating form was developed by one of the author's (AJM) to collect faculty ratings of five broadly defined noncognitive skills of clinical dental students (Appendix A). The direction of scoring was reversed on alternating scales to minimize a response set of marking the same point on each scale. For phase 1 of the study, 32 faculty were asked to rate 31 dental students at the end of their junior year. Twenty-nine faculty complied; however, only 12 evaluated all 31 students on each of the five scales (i.e.,

did not make use of the "Not Applicable" rating). For phase 2 of the study, 29 of the 31 students were evaluated at midsemester of their senior year by 9 different faculty. Complete ratings were available from 5 raters for 16 students.

A standard factorial analysis of variance program (BMD 02V; Dixon, 1973) was utilized to compute the various mean squares utilized in the study. Computational formulas derived by Cardinet et. al., (1976) were utilized in computing the generalizability coefficients.

Results and Discussion

Table I summarizes the mean scores and standard deviations across students for each scale.

Table I

Descriptive Statistics

	Scale I	Scale 2	Scale 3	Scale 4	Scale 5
Phase I (Jr. Year) n = 31 X 12	$\bar{X}_{11} = 3.41$	$\bar{X}_{12} = 3.53$	$\bar{X}_{13} = 3.35$	$\bar{X}_{14} = 3.43$	$\bar{X}_{15} = 3.71$
	$\sigma_{11} = 1.17$	$\sigma_{12} = 1.19$	$\sigma_{13} = 1.13$	$\sigma_{14} = 1.10$	$\sigma_{15} = 1.09$
Phase 2 (Sr. Year) n = 16 X 5	$\bar{X}_{21} = 3.75$	$\bar{X}_{22} = 3.88$	$\bar{X}_{23} = 3.61$	$\bar{X}_{24} = 3.74$	$\bar{X}_{25} = 4.11$
	$\sigma_{21} = 1.09$	$\sigma_{22} = 1.23$	$\sigma_{23} = 1.07$	$\sigma_{24} = 1.03$	$\sigma_{25} = 1.02$

Table II describes three potential uses of this measurement system, provides a computational model for each potential use, and displays the resulting generalizability coefficients based upon data from phases 1 and 2 of the study.

Table II

Computing Generalizability Coefficients

Purpose of Measurement System	Computational Formula	Generalizability Coefficient Phase 1 (N = 31X12)	Generalizability Coefficient Phase 2 (N = 16X5)
1. Differentiate among students using five given scale scores regardless of which faculty provide the ratings.	$\rho^2_{St} = \dfrac{\sigma^2_{St} + \sigma^2_{St \cdot Sc}/N_{Sc}}{\sigma^2_{St} + \sigma^2_{St \cdot Sc}/N_{Sc} + \sigma^2_{St \cdot R}/N_R + \sigma^2_{St \cdot Sc \cdot R}/N_R N_{Sc}}$	= 0.95	= 0.86
2. Differentiate among students using a mean score across scales without regard for which faculty provide the ratings.	$\rho^2_{St} = \dfrac{\sigma^2_{St}}{\sigma^2_{St} + \sigma^2_{St \cdot Sc}/N_{Sc} + \sigma^2_{St \cdot R}/N_R + \sigma^2_{St \cdot Sc \cdot R}/N_R N_{Sc}}$	= 0.92	= 0.83
3. Assess the inter-scale differences for a given student.	$\rho^2_{Sc} = \dfrac{\sigma^2_{Sc}}{\sigma^2_{Sc} + \sigma^2_{St \cdot Sc}/N_{St} + \sigma^2_{Sc \cdot R}/N_R + \sigma^2_{St \cdot Sc \cdot R}/N_{St} N_R}$	= 0.75	= 0.62

N_R = # raters N_{St} = # students N_{Sc} = # scales

The coefficients in Table II indicate that this measurement system can adequately differentiate among dental students using five scale scores (model 1) or using a mean score across scales (model 2). Model 2 is commonly used in research studies related to the inter-rater reliability of rating scales (Ebel, 1951; Gaines et. al., 1975; Rowley, 1976). The coefficients based upon junior year ratings (phase 1) are larger than the coefficients based upon the senior year ratings (phase 2). These differences are no doubt a reflection of the fact that for phase 2, complete data were available for only 16 students rated by 5 senior-year faculty. The generalizability coefficients for model 3 indicate that one should be careful when interpreting differences among rating scale scores for a given student.

An additional feature of generalizability theory is the capability of estimating the changes in generalizability coefficients to be expected if

the measurement conditions are altered. For example, Table III demonstrates the generalizability coefficients which would be expected if the number of observers is reduced to 1 or 2.

Table III

Effects of Reducing the Number of Raters

	Phase 1		Phase 2	
	$N_R = 2$	$N_R = 1$	$N_R = 2$	$N_R = 1$
Model 1*	$\rho^2_{st} = .76$	$\rho^2_{st} = .61$	$\rho^2_{st} = .71$	$\rho^2_{st} = .55$
Model 2*	$\rho^2_{st} = .75$	$\rho^2_{st} = .60$	$\rho^2_{st} = .68$	$\rho^2_{st} = .53$
Model 3*	$\rho^2_{sc} = .49$	$\rho^2_{sc} = .34$	$\rho^2_{sc} = .45$	$\rho^2_{sc} = .31$

*Model numbers correspond to Table II

The coefficients in Table III provide an objective basis for making decisions regarding the minimum number of observers required to maintain prespecified levels of generalizability (e.g., .75). It is apparent that with only one observer, interpretations of inter-scale differences for a given student (model 3) must be quite tentative (i.e., $\rho^2 = .31 - .34$).

Conclusions

A clinical rating instrument is useful only to the extent that it accurately quantifies information regarding the existence of a student trait or characteristic. Various types of instructional decisions may then be made depending upon the precision of the measurement information. In actual practice. important decisions are often made even though the "numbers" might have been very different if the measurement conditions had been altered (e.g., another observer had been involved).

Generalizability theory offers several advantages over classical test theory as a model for assessing the precision of the "numbers" generated by a particular measurement system (in this case, a five scale rating form with multiple faculty observers). The major implication of the study is that when clinical rating instruments are developed, the pilot testing should include computation of a generalizability coefficient for each intended use of the scores (e.g., interpreting scale scores of an individual student). Implementation decisions such as determining the number of observers or the number of student trials can then be based upon objective data in relation to maintaining a certain minimum generalizability coefficient. By way of illustration, the rating instrument used in this study would require at least 2 dental faculty raters to maintain a coefficient of approximately .70 when differentiating among students.

Bibliography

Cardinet, J., Tourneur, Y. and Allal, L. The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 1976, 13, 119-135

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972

Dixon, W.J. BMD Biomedical computer programs. Los Angeles: University of California Press, 1973

Ebel, R.L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424

Gaines, W.G., Bruggers, H. and Rasmussen, R.H. Reliability of the ratings of products of dental morphology. Paper presented at the annual meeting of the American Educational Research Association, Washington, April, 1975

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950

Rowley, G.L. The reliability of observational measures. American Educational Research Journal, 1976, 13, 51-59

INSTRUCTIONS TO FACULTY: Please rate the student on the following characteristics by placing an X in one of the appropriate boxes. Rate each characteristic independently of the ratings given on previous characteristics. Give your own opinion. Do not consult with your colleagues

1. *CONCEPT OF TREATMENT: Has a good concept of treatment to be performed as well as a good concept of the total dental needs of patients.*

	5	4	3	2	1	
excellent concept	☐	☐	☐	☐	☐	*poor concept*

☐ *I feel I do not know this student well enough to rate him/her on this characteristic.*

2. *FOLLOWS SUGGESTIONS: Accepts constructive criticism and applies it effectively.*

	1	2	3	4	5	
poorly	☐	☐	☐	☐	☐	*excellently*

☐ *I feel I do not know this student well enough to rate him/her on this charactertistic.*

3. *SELF-CONFIDENCE & RELIANCE: Has self-confidence, takes initiative, and performs correct treatment independently.*

	5	4	3	2	1	
excellent self-confidence	☐	☐	☐	☐	☐	*poor self-confidence*

☐ *I feel I do not know this student well enough to rate him/her on this characteristic.*

4. *ORGANIZATION & PREPARATION: Organized and prepared to perform treatment, operatory is neat. Instruments, materials, and records are ready when instructor arrives.*

	1	2	3	4	5	
poorly organized	☐	☐	☐	☐	☐	*well organized*

☐ *I feel I do not know this student well enough to rate him/her on this characteristic.*

5. *EMPATHY: Demonstrates empathy and concern for patients' needs and comfort.*

	5	4	3	2	1	
excellent	☐	☐	☐	☐	☐	*poor*

☐ *I feel I do not know this student well enough to rate him/her on this characteristic.*